Scalability analysis methodology for passive optical interconnects in data center networks using PAM

Scalability Analysis Methodology for Passive Optical Interconnects in Data Center Networks Using PAM

R. Lin[a,b], Krzysztof Szczerba[c], Erik Agrell[d], Lena Wosinska[b], M. Tang[a], D. Liu[a], J. Chen[b].

[a] *Next Generation Internet Access National Engineering Lab (NGIA), School of optical and electronic information, Huazhong University of Sci&Tech (HUST), 1037 Luoyu Road, Wuhan, China.* (*e-mail:rulin@kth.se, tangming@mail.hust.edu.cn, dmliu@mail.hust.edu.cn)*

[b] *School of Information and Communication Technology, KTH Royal Institute of Technology, Electrum 229, Kista, Sweden. (wosinska@kth.se, jiajiac@kth.se)*

[c] *Department of Microtechnology and Nanoscience, Chalmers University of Technology, Gothenburg, Sweden, now at now at Finisar Corp., Sunnyvale, CA, USA (krzysztof.szczerba@finisar.com)*

[d] *Department of Electrical Engineering, Chalmers University of Technology, Göteborg, Sweden. (agrell@chalmers.se)*

Abstract

A framework is developed for modeling the fundamental impairments in optical datacenter interconnects, i.e., the power loss and the receiver noises. This framework makes it possible, to analyze the trade-offs between data rates, modulation order, and number of ports that can be supported in optical interconnect architectures, while guaranteeing that the required signal-to-noise ratios are satisfied. To the best of our knowledge, this important assessment methodology is not yet available. As a case study, the trade-offs are investigated for three coupler-based top-of-rack interconnect architectures, which suffer from serious insertion loss. The results show that using single-port transceivers with 10 GHz bandwidth, avalanche photodiode detectors, and quadratical pulse amplitude modulation, more than 500 ports can be supported.

Keywords: *Data center networks, passive optical interconnect (POI), multi-level pulse amplitude modulation (M-PAM), scalability, top-of-rack (ToR).*

I. Introduction

The global data center (DC) traffic is ever growing [1]. It is expected that the total DC traffic will reach 8.6 ZB per year by 2018 and that 75% of this traffic will stay within the DCs [2]. The current DC networks (DCNs) based on commodity Ethernet switches suffer from the bandwidth bottleneck and high energy consumption [3]. As the size and the complexity of DCs continue to grow, scaling out of the current DC infrastructure becomes challenging. Therefore, the optical transport technology, which has been widely adopted in telecommunication networks for its high capacity, short latency, and low power consumption, can be a prominent option for DCNs as well [4, Part I]. Typically, intra-DCN communications includes two to three tiers [5], i.e., edge tier, aggregation and/or core tier, where switches handle the traffic among the servers, racks and/or clusters, respectively. Many interconnect architectures based on optical switching for DCNs have already been studied. Hybrid optical/electrical solutions, e.g., HELIOS [6] and c-through [7] have been proposed to increase the capacity and reduce the power consumption. The main concern with hybrid approaches is the inherent bottleneck in terms of energy dissipation of using electrical switches to handle the traffic at the packet level. All-optical interconnect architectures at the aggregation and core tiers have also been proposed, e.g., [8]–[10]. However, they suffer from the high cost of the non-commercial devices [8], [9] and high end-to-end latency due to the long configuration time of the switches [10].

Apart from the aforementioned optical switching based solutions, passive optical interconnects (POIs) have also been proposed, e.g., in [11]–[13]. Passive components, such as couplers, are used to interconnect different servers, e.g., in [11], [12] and racks/clusters, e.g., in [13]. Passive optical networks (PONs) [14] that are widely deployed in fiber access networks use optical passive components, such as splitters/combiners, to connect the central office to multiple end users. However, it cannot be directly applied to datacenter network where connections between any pair of ports are required. There are many advantages of coupler-based POIs, such as low cost and energy consumption, high reliability, simple maintenance and ability to support multi-cast and any-cast connections [15]. It has been demonstrated that the power consumption of switching in overall DCNs can be reduced by a factor of 10 for large-scale DCs when POIs are employed at the edge tier [16]. In contrast to switch-based optical interconnect, POIs seriously suffer from physical-layer impairments, such as high insertion loss caused by the couplers, resulting in limited scalability. Such a scalability problem in switch-based architectures is addressed by scaling the size of the switch. There are also works measuring scalability from the aspects of the control and management plane [17], [18]. However, the aforementioned scalability assessment methods are not applicable for POI architectures where the physical layer brings the main scalability limitation. In [12] and [13], scalability analysis of certain POI architectures is performed by comparing the maximum link loss and a given system power budget. Apparently, considering only the link loss is not sufficient, because the other physical-layer configuration, e.g., the type of receivers and modulation formats, can also significantly affect the scalability.

In this paper, we develop a physical-layer evaluation methodology, which is vital to analyze scalability of the POI architectures. The proposed methodology enables to investigate the trade-offs between data rates, modulation formats, and number of ports that can be supported by optical interconnects. We focus on intensity modulation and direct detection (IM/DD), which is considered as one of the best options for DC applications, because it is simple and cost-efficient. For IM/DD techniques, on-off keying (OOK) is already widely used, whereas multi-level pulse amplitude (M-PAM) modulation is currently of high interest for datacom applications thanks to its low implementation complexity and cost to increase spectrum efficiency and upgrade capacity [19]. The influence of PAM symbol spacing and decision thresholds is also included in the proposed methodology. The scalability of the optical interconnect employing equally and optimally spaced M-PAM is compared. Moreover, two major types of photodiodes (PDs) at the receiver, i.e., positive-intrinsic-negative (PIN) diode and avalanche photodiode (APD) are taken into account for modeling receiver sensitivity. A case study is carried out, where the proposed methodology is applied to assess three coupler-based POIs. The analysis results reveal that with a proper physical-layer settings (including modulation format, receiver type, etc.), more than 500 ports and 10Gb/s per port can be supported, even in the case of coupler-based POIs with high insertion loss.

The remainder of this paper is organized as follows. In Section II, the proposed physical-layer scalability assessment methodology is presented. Section III first depicts three investigated scenarios for the case study, each of which employs one type of coupler-based POI at the edge tier of the DCN (i.e., at ToR), and then applies the proposed methodology to identify the maximal data rates and number of ports that can be supported in all three considered scenarios. Finally, Section IV draws the conclusions.

II. Methodology

The scalability of the conventional switch-based interconnect architectures is limited by the port count of the switch [8–10], which is not the case for POIs. On the other hand, the scalability assessment in [12] and [13], considers only the insertion loss of the couplers. In this section, we introduce a methodology to assess the scalability of optical interconnects, where different physical layer aspects are considered. The flow chart of our scalability assessment methodology is shown in Fig. 1. We build the link model to analyze the link loss. Meanwhile, the system power budget is obtained by assessing the receiver sensitivity where the type of receiver, the M-PAM order and spacing, the decision thresholds, and the allowed bit error rate (BER) level are considered. If the system power budget is larger than the link budget, the interconnect with the considered size and the data rate at each port are demonstrated to be feasible. The details for modeling link loss and system power budget are presented in the following subsections.

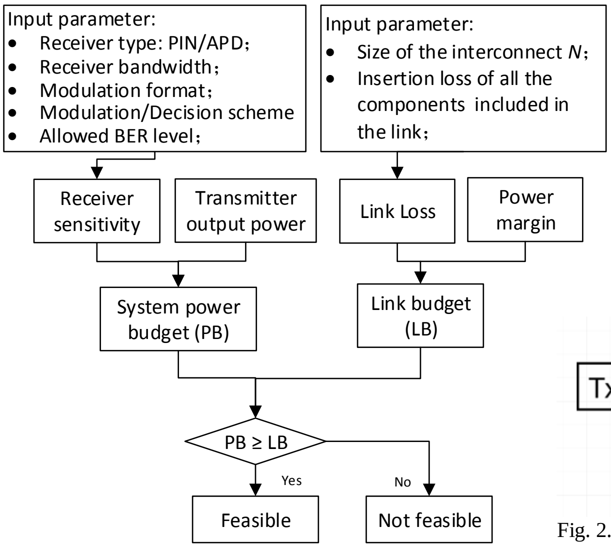


Fig. 1. Flow chart of the scalability assessment methodolc

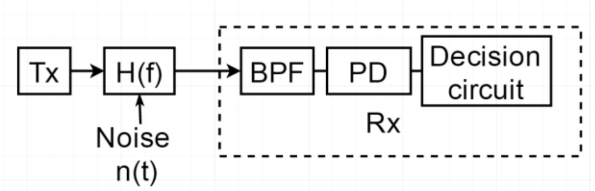


Fig. 2. Equivalent block diagram for the optical interconnect link. Tx: transmitter, BPF: band pass filter, PD: photo diode, Rx: receiver.

A. Link budget estimation

For intra-DC applications, the transmission distance is typically quite short. For instance, the intra-rack signal transmission distance is normally shorter than 10 meters (between bottom and top of the rack). Also, we consider the case without any optical amplifiers, so amplified spontaneous emission noise does not exist. Fig. 2 shows an equivalent block diagram for the optical interconnect link.

In this study, we focus on single-mode fibers (SMF), which operate at 1550 nm. While multimode fiber (MMF) for 850 nm is commonly used for short range communications today, SMF technology allows dense-wavelength division multiplexing (DWDM) technology, offering high capacity and interoperability throughout the intra-DC and inter-DC networks [20]. In addition, the stringent temperature control, which is required by DWDM devices, is inherently provided by the indoor environment of DCs. In the absence of fiber nonlinearity, the link transfer function of an SMF with length L can be written as [26]

$$H(f) = A \cdot exp\left(j\pi \frac{\lambda^2}{c} LDf^2\right), \tag{1}$$

where A, λ, c, D, and f denote the attenuation caused by the link, the wavelength, the speed of the light, the chromatic dispersion parameter at λ, and the signal frequency, respectively. Given the scenario, where λ, c, D and L have the values of 1550 nm, 3×10^8 m/s, 1.6 ps/nm·km, and 10 m, respectively, if the signal frequency $|f| \leq 100$ GHz, (1) is well approximated by

$$H(f) \approx A. \tag{2}$$

Hence, the chromatic dispersion can be neglected. A similar calculation shows that fiber nonlinearity is also negligible at these short distances.

The total link loss stands for the optical power attenuation caused by passing all the components in the link. The link budget can be expressed as

$$Link\ Budget = IL_{total} + P_{margin} = \sum_{i=0}^{n} IL_i + P_{margin}, \tag{3}$$

where IL_{total} represents the total insertion loss in the link ($IL_{\text{total}} = 20\log_{10}A$ in dB), IL_i stands for the insertion loss of each component, and P_{margin} is the power margin that is typically reserved for other impairments, such as insertion loss caused by connectors and splices. In POI-based architectures, e.g., those presented in [13] and [21], the use of optical couplers leads to a high link loss. The splitting loss sets a lower bound on the coupler insertion loss [27, Ch. 8]. The commercial couplers usually have port count as exponents of 2, e.g., 8×1, 16×2. The minimum loss due to passing the coupler once is $10\log_{10}n$ in dB, where n is the coupler port count (for the case of $n\times m$ coupler, the port count is n if n ≥ m and m otherwise).

B. System power budget assessment

The system power budget can be calculated based on the difference between the launch power of the transmitter and the receiver sensitivity. The launch power is typically a given input parameter of the transmitter. At the receiver side, the band-pass filter (BPF) filters out the wavelength of interest and then the PD converts the terminated optical power to a photocurrent. The decision circuit determines the most probable value of the input symbol based on the received power. We assume that the BPF has a flat frequency response. How to model the receiver noise depends on the type of receiver. Here we consider both a PIN diode and an APD.

The photocurrent converted by the PIN diode is

$$I_{\mathrm{opt}} = R_{\mathrm{d}} \cdot P_{\mathrm{opt}}, \tag{4}$$

where R_{d} is the responsivity of the photodiode and P_{opt} is the received optical power. The total noise current variance in the receiver is

$$\begin{aligned} \sigma^2 &= \sigma_T^2 + \sigma_S^2 + \sigma_{RIN}^2 \\ &= 4k_B T F_n \Delta f / R_L + 2q(I_{opt} + I_d)\Delta f + RIN \cdot I_{opt}^2 \Delta f \end{aligned} \tag{5}$$

where the three terms on the right side of (5) represent thermal noise, shot noise, and relative intensity noise (RIN), respectively [32, Ch. 5]. The current fluctuation induced by thermal noise is mathematically modeled as a stationary Gaussian random process. In thermal noise, k_B, T, F_n, Δf, and R_{L} are the Bolzmann constant, the temperature in Kelvin, the noise figure of the receiver amplifier, the bandwidth of the photodiode, and the resistance, respectively [32, Ch. 5]. The shot noise induced current variance is proportional to the combination of incident current I_{opt} and dark current I_{d} where q is the elementary charge. I_{d} is usually far smaller than I_{opt} and can be neglected. The RIN depends on the square of the photocurrent where RIN is the average relative intensity noise spectral density.

Optical receivers that employ APDs generally provide better sensitivity than the ones with PIN diodes. This improvement is due to the internal gain in an APD that increases the photocurrent by a multiplication factor M_{APD} [32, Ch. 5.2] so that

$$I_{\mathrm{opt}} = M_{\mathrm{APD}} R_d P_{\mathrm{opt}} \,. \tag{6}$$

Thermal noise and RIN in an APD can be also expressed as (5). The shot noise of an APD can be expressed as a function of M_{APD} and the ionization coefficient k_A [30, Ch. 5.2] as

$$\sigma_s^2 = 2q\, M_{\mathrm{APD}}^2 \left[k_{\mathrm{A}} M_{\mathrm{APD}} + (1 - k_{\mathrm{A}})\left(2 - \frac{1}{M_{\mathrm{APD}}}\right)\right] R_{\mathrm{d}} P_{\mathrm{opt}} \Delta f. \tag{7}$$

Assuming additive white Gaussian noise [32, Ch. 3], the total current noise power is proportional to the photodiode bandwidth. The receiver sensitivity, defined as the minimum average power required by the receiver to reach a certain level of BER, will be affected by the different noise current at different PDs, and the modulation scheme should be adjusted accordingly for better sensitivity.

Using M-PAM with Gray labeling, the BER can be approximated as a function of the symbol error rate (SER) [17]

$$BER \approx SER / \log_2 M = \frac{1}{M} \sum_{i=0}^{M-1} \sum_{j=0, j \neq i}^{M-1} P_{ij} \tag{8}$$

P_{ij} is the probability of transmitting PAM symbol i but receiving symbol j. It can be expressed as [33, Ch. 5]

$$P_{ij} = \frac{1}{2}\operatorname{erfc}\left(\frac{I_{th,j} - I_i}{\sigma_i\sqrt{2}}\right) - \frac{1}{2}\operatorname{erfc}\left(\frac{I_{th,j+1} - I_i}{\sigma_i\sqrt{2}}\right), \tag{9}$$

where I_i is the photocurrent of symbol i, and its average value satisfies $\left(\frac{1}{M}\right)\sum_i I_i = I_{\mathrm{opt}}$. $I_{th,j}$ denotes the photocurrent threshold between symbols j and j-1, and σ_i is the noise current variance at symbol i and can be calculated using (5)–(9). $I_{th,0}$ should be interpreted as $-\infty$ and $I_{th,\,M}$ as $+\infty$. Assuming that errors occur only between adjacent symbols when the adjacent symbols contribute equally to the symbol error probability, the optimal threshold level $I_{th,i}$ that minimizes the error rate [32, Ch. 5.3] can be obtained by:

$$I_{th,i} = \frac{I_i \sigma_{i-1} - I_{i-1}\sigma_i}{\sigma_{i-1} + \sigma_i}. \tag{10}$$

The minimum BER can be achieved with equispaced symbols and decision thresholds halfway between the symbols when thermal noise dominates, because it is independent of the optical system level, i.e., $\sigma_i = \sigma_{i-1}$. However, the equispaced modulation is suboptimal [33, Ch. 5] when shot noise, which is linearly proportional to the received optical power, is the main factor in the overall noise. Instead, quadratically spaced symbols in modulation is found to be optimum [34], where the normalized weight for the ith symbol is $[(i-1)/(M-1)]^2$. The corresponding optimal decision threshold can be obtained using (10). The receiver sensitivity, i.e., the minimum received optical power that satisfies a given BER requirement, can be obtained by (4)–(10) for various types of receivers, available bandwidth at the receivers, and modulation formats. The system power budget can be finally calculated according to the difference between the launch power of the transmitter and the receiver sensitivity. The optical interconnect architecture with size N is feasible if the link budget is lower than the system power budget.

III. Case Study

In this section, we first present three coupler-based POI architectures for a case study and then apply the proposed methodology to evaluate their scalability.

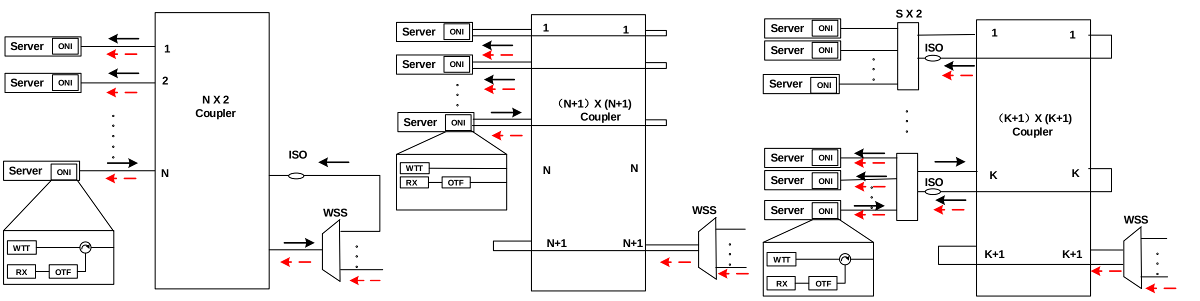
A. *Coupler-based POIs*

Fig. 3 illustrates the principle of operation of the considered interconnects at ToR. Each server in the rack has one optical network interface (ONI) that can transmit and receive one wavelength at a time. DWDM technology can be employed to fully utilize the optical spectrum. A wavelength-tunable transmitter (WTT) and an optical tunable filter (OTF) are adopted at each ONI for wavelength tunability. Tuning the transceivers can be realized in nanoseconds [22]–[24]. Meanwhile, fast tunable filters [25] are technically feasible and have the potential to be commercialized. The price of such transceivers can be significantly reduced once the market demand increases.

The spectrum is divided into two sets of wavelengths [11]. One set is used for traffic within the rack, while the other set is allocated for inter-rack communication. If these POIs are applied in the aggregation or core tier, one set of wavelengths handles traffic among different racks or clusters and the other set is needed for traffic passing the interfaces to the upper tier or to the Internet. Furthermore, to avoid contention at the receiver side, time division multiplexing access combined with DWDM needs to be applied to coordinate transmission from several servers to the same destination. An appropriate MAC protocol, e.g., the one proposed in [13], can be applied in all three considered POI architectures. We consider N servers per rack (or N racks/clusters at the aggregation/core tier). Here, N represents the size of the interconnect. In all presented POI architectures, a DWDM spacing of 50 GHz and signals of 10 GBaud are used, so that crosstalk can be neglected.

1) *Scheme I*

Scheme I is an interconnect architecture using an N×2 coupler (see Fig. 3(a)). It was originally proposed in [11] as the interconnect at the ToR. ONIs at the N servers are connected to the input ports of the N×2 coupler via SMFs. Each ONI is equipped with a transceiver with a single fiber port. Taking Server N as an example, the intra-rack signals transmitted from Server N (Fig. 3(a), solid black arrows) are broadcast to all the servers (including Server N itself) in the rack. This is because the wavelengths assigned for intra-rack communication are switched to the output port of the WSS that is connected to the coupler through an isolator inserted between the coupler and the WSS to keep the light transmission in one direction only and avoid self-interference. The remaining WSS ports are serving the signals to and from outside of the rack. Here, the WSS reserves several ports towards the upper tier, which are often required for high capacity and resilience. The dashed red arrows show the traffic received from the outside of the rack, which is broadcast to all the servers in the rack. The OTF at the ONI filters out the wavelength assigned to this server. The intra-rack signal suffers from a large power loss due to passing the coupler twice.



(a) Scheme I (b) Scheme II (c) Scheme III

Fig. 3. Architectures of coupler-based POI. (ONI: optical network interface; WTT: wavelength tunable transmitter; RX: receiver; OTF: optical tunable filter; WSS: wavelength selective switch; ISO: isolator.)

2) *Scheme II*

In order to reduce the insertion loss, Scheme II (see Fig. 3(b)) was recently proposed in [20]. It contains an (N+1)×(N+1) coupler. Each ONI in Scheme II needs two fiber ports connected to the coupler, i.e., one to transmit and one to receive the optical signals. A dual-port transceiver is thus adopted in each ONI. Again taking Server N as an example, the intra-rack signals transmitted from Server N are broadcast to all the servers in the rack after passing the (N+1)×(N+1) coupler (solid black arrows in Fig. 3(b)). The signals to and from the other racks are selected and directed by the WSS. In Scheme II, the WSS needs two ports connected to the coupler. On the other hand, similar to Scheme I, the WSS reserves several ports towards the upper tier. The dashed red arrows in Scheme II represent the traffic received from the outside of the rack, which is broadcast to all the servers in the rack after passing the coupler. Since the intra-rack signals only pass the coupler once, the insertion loss is reduced significantly compared to Scheme I. Besides, the use of two fiber ports for each ONI maintains unidirectional transmission in all the fibers. Consequently, it eliminates the need of an isolator and a circulator at each ONI.

3) *Scheme III*

Scheme III (see Fig. 3(c)) is proposed as a compromise between Scheme I and Scheme II in terms of insertion loss and cabling complexity. It needs two stages of relatively small couplers, compared to the ones required in Schemes I and II, where a single but large-size coupler is necessary. An isolator is required to connect the two stages of couplers in order to keep the unidirectional transmission of light. In the first stage, a set of K S×2 couplers are employed, while a (K+1)×(K+1) coupler is used in the second stage, where $K = \lceil N/S \rceil$ and the ceiling function $\lceil x \rceil$ represents the smallest integer that is not less than x. In this scheme, all the servers are equipped with single-port ONIs. The intra-rack data transmitted from the servers passes the first-stage coupler twice and the second-stage coupler once to be broadcast to all the servers in the rack (see, for example, the solid black arrows in Fig. 3(c), which represent the traffic from Server N). With the same server count N, when the number of input ports at the first stage coupler, i.e., S, is doubled, the insertion loss is in principle increased by at least 3 dB. Therefore, from the insertion loss point of view, it is good to keep the size of first stage coupler as small as possible. A 2×2 coupler is the smallest size that can be used at the first stage. When S = 2, the overall insertion loss for intra-rack signals caused by the coupler in Scheme III is increased by approximately 3 dB compared to the one in Scheme II. On the other hand, it can still have a significant reduction of intra-rack link loss compared to Scheme I if the number of ports N of the interconnect is large (e.g., $N \geq 16$). In Scheme III, the size of couplers in the first and second stages can be flexibly selected.

Among the three architectures, the link in Scheme II has the least power loss as the signals only pass the coupler once and do not need to go through any isolators or circulators. On the other hand, Scheme I and III just require single-port transceivers while Scheme II needs dual-port ones to keep unidirectional propagation in the connected fibers. Dual-port transceivers may increase the cabling complexity and the required footprint.

TABLE I
PARAMETER TABLE [29, CH. 2]

Symbol	QUANTITY
T	304 K
F_n	5 dB
Δf	10 GHz
R_L	50 Ω
R_d	1 A/W
RIN	−145 dB/Hz
M_{APD}	10
k_A	0.45

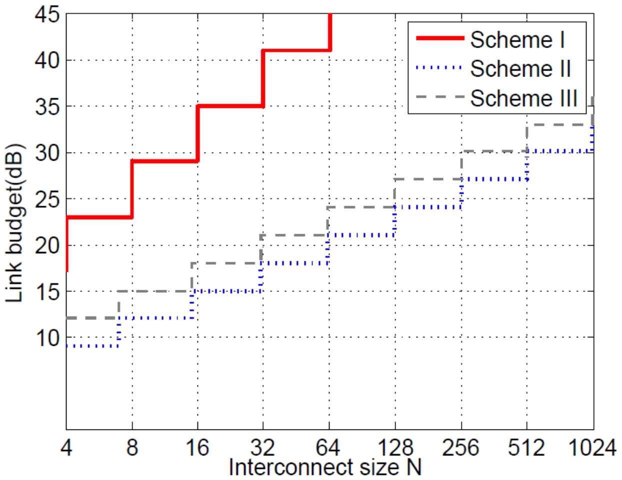


Fig. 4. The optical link budget for communication within the interconnect as a function of the interconnect size N.

B. *Scalability analysis*

In this section, we apply the methodology proposed in Section II to the considered interconnect architectures. The link budget is calculated and the system power budget is assessed. The scalability analysis of each architecture is performed by comparing these two values.

In all three coupler-based POI architectures, the total insertion loss can be expressed as

$$IL_{total} = IL_{coupler} + IL_{WSS} + IL_{ISO} + IL_{CIR} + IL_{OTF}, \tag{16}$$

where $IL_{coupler}$, IL_{WSS}, IL_{ISO}, IL_{CIR}, and IL_{OTF} refer to the insertion loss of the coupler(s), WSS, isolator, circulator, and OTF, respectively. Note that the insertion loss calculation of intra-rack connections in Scheme II does not include the WSS and isolator. The NxN coupler can be built by cascading several small ones, e.g., 2x2 or 3x3 couplers, which may introduce additional losses [13]. In this calculation, we consider the lower bound of the coupler insertion loss. ILcoupler in all the three schemes is different. Intra-rack signals pass the coupler twice in Scheme I leading to the coupler loss in dB:

$$IL_{\text{coupler}} = 6\,\lceil \log_2 N \rceil. \tag{17}$$

In Scheme II, the intra-rack signals pass the coupler only once. Thus, the total coupler loss can be written as

$$IL_{\text{coupler}} = 3\,\lceil \log_2 (N+1) \rceil. \tag{18}$$

In Scheme III, the signals pass the first stage coupler twice and the second stage once, so the coupler splitting loss is

$$IL_{\text{coupler}} = 6\,\lceil \log_2 S \rceil + 3\,\lceil \log_2 (K+1) \rceil. \tag{19}$$

In the link budget calculation, the values of IL_{WSS}, IL_{ISO}, and IL_{CIR} are 2 dB [28], 0.4 dB [29], and 0.6 dB [30], respectively, according to the datasheets of commercially available products. We assume IL_{OTF} = 0 dB if employing the technique proposed in [31]. Apart from the insertion loss of the components, 2 dB margin is reserved for compensating impairment penalties and connection loss in the link. The relationship between the number of the servers per rack, i.e., the interconnect size N, and the link budget can be found in Fig. 4. We consider using a 2×2 coupler in the first stage in Scheme III, showing the lower bound of its link budget. Scheme I is characterized by the largest link budget among the three optical interconnect architectures studied in this paper. It is more than 10 dB larger than that of Scheme II, which offers the smallest link budget.

To support high capacity DC traffic requirement, a 10 GHz transceiver with PIN diode or APD is assumed for each server. With the parameters given in Table I, the noise contribution in both PIN diode and APD receivers are obtained and shown in Figs. 5 and 6 respectively. In the receiver with a PIN diode, the dominant noise is the thermal noise, which means that the minimum BER is achieved with equispaced symbols and decision thresholds halfway between the symbols. The theoretical BER performance of M-PAM as a function of the received optical power is shown in Figs. 7 and 8. Four cases are calculated, namely, 2-PAM (OOK), 4-PAM, 8-PAM, and 16-PAM. With full use of the 10 GHz PIN diode, the data rate of M-PAM is $\log_2$ M×10 Gbps. For error-free transmission, the receiver sensitivity of a 10 GBaud OOK signal is around –16 dBm. There is about 4.8 dB sensitivity degradation when the number of symbols per bit is doubled.

An error floor can be observed when 16-PAM is used. That is because the RIN dominates the noise when the received power is over –4 dBm, as shown in Fig. 7. When an APD is employed at

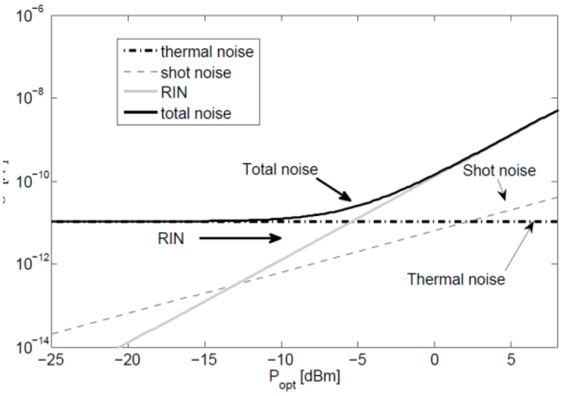


Fig. 5. Thermal noise, shot noise, relative intensity noise, and total noise for the given system with a 10 GHz bandwidth PIN diode at the receiver.

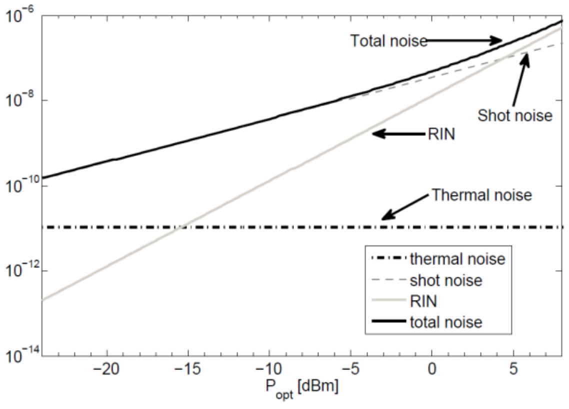


Fig. 6. Thermal noise, shot noise, relative intensity noise, and total noise for the given system with a 10 GHz bandwidth APD at the receiver.

the receiver, shot noise contributes most to the total noise. In this case, equal symbol spacing is suboptimal while quadratically spaced symbols are optimum [34], where the normalized weight for the ith symbol is $[(\boldsymbol{i}-\mathbf{1})/(\boldsymbol{M}-\mathbf{1})]^{\mathbf{2}}$. By using different forward error correction (FEC) techniques the BER can be brought down to $\mathbf{10^{-12}}$ or less after decoding. The corresponding threshold is obtained using (10). An improvement of 5.4 dB and 6.8 dB in sensitivity can be obtained with the optimal symbol spacing for 10 GBaud 4-PAM and 8-PAM, respectively (see Fig. 8).

With the sensitivity calculated, the power budget can be obtained for each architecture with different modulation formats, as the difference between the launch power of the transmitter and the receiver sensitivity. The higher the launch power is, the larger the available power budget can be achieved. Here we assume that the transmitter launch power is 10 dBm [35], which is a typical value for a commercial transmitter. For each scheme with a given interconnect size N, we try to find the power budget with the highest order of modulation format, representing the highest data rate that can satisfy the link budget calculated in Section III. For instance, with 8 servers per rack, the link budget of Scheme I is equal to 23 dB, which is obtained by adding 2 dB margin to the link loss in (3). To realize error- free transmission (i.e., BER = 10^{-12}), only 4-PAM and OOK provide the required power budget, which is larger than 23 dB, no matter which receiver is used. Thus, in this case, the highest data rate that Scheme I can handle is 20 Gbps, achieved by 4-PAM. When the number of servers per rack in Scheme I is higher than 8, neither the APD nor the PIN diode is able to support a BER lower than 10^{-12}. Therefore, in order to obtain error-free transmission, the maximum size of interconnect (i.e., the maximum number of servers in a rack) in Scheme I is 8.

Fig. 9 shows the maximum data rate as a function of the rack size that can support error-free transmission in each scheme. We consider the first stage couplers in Scheme III as a set of 2×2 couplers, i.e., S = 2. The maximum number of ports in the coupler-based optical interconnects is 255. It can be realized by both the PIN diode and the APD receiver. On the other hand, neither the PIN diode nor the APD can support 40 Gbps error-free transmission, i.e., 16-PAM cannot be supported. To achieve the maximum data rate of 30 Gpbs, the size of coupler-based interconnect should not exceed 15. The PIN diode and the APD have similar performance when the interconnect size is in the range of 4 to 63. The APD provides better scalability for error-free transmission when the size is larger than 63. For Scheme III, when S is doubled, the power loss increases by 3dB at least, and the maximum size of the interconnect operating at the same data rate reduces at least by half.

The scalability of all the schemes can be significantly improved when FEC is implemented, i.e., the receiver sensitivity is measured at the FEC threshold of BER = 10^{-3} (see Fig. 10). In case where a PIN diode is used, the maximum size of the coupler-based optical interconnects is 511, and 40 Gbps data rate can be handled with no more than 31 servers in a rack. The APD provides even better scalability for all the schemes. Up to 2047 ports can be served using APD receivers. It is because the APD has better BER performance than the PIN diode when the received optical power is low (< −10 dBm). Obviously, Scheme I is the least scalable one in terms of the number of ports, which limits the interconnect size that can be supported. Schemes II and III demonstrate significant advantages in both rack size and maximum data rate. Particularly, Scheme II always

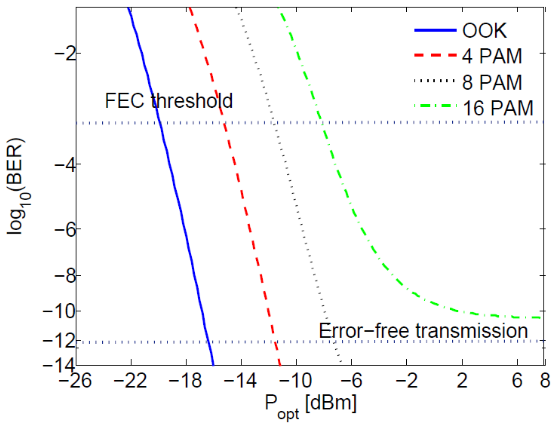


Fig. 7. BER performance of 10 GBaud *M*-PAM with a PIN diode at the receiver.

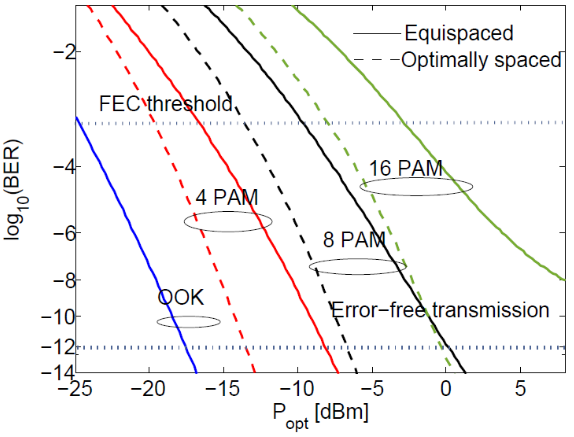


Fig. 8. BER performance of 10 GBaud *M*-PAM with an APD at the receiver using equidistant (solid lines) and optimal symbol spacing (dashed lines).

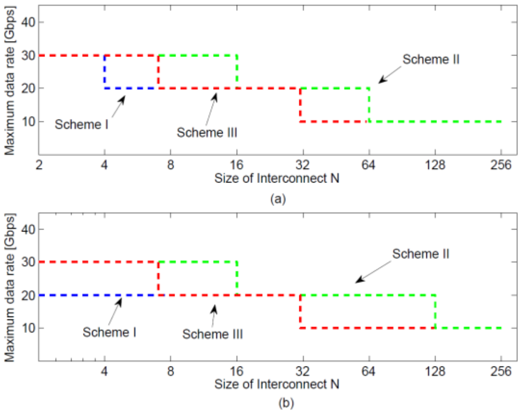


Fig. 9. The relationship between the interconnect size *N* and the maximum data rate in error-free transmission (BER= 10^{-12}) with a receiver employing (a) a PIN diode and (b) an APD.

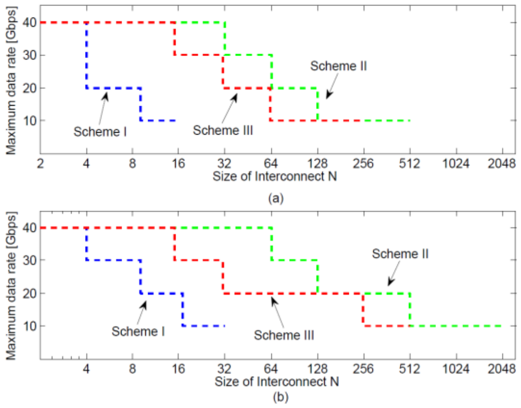


Fig. 10. The relationship between the interconnect size *N* and the maximum data rate under the FEC threshold of BER=10^{-3} with a receiver employing (a) a PIN diode and (b) an APD.

supports the highest maximum data rate and the largest interconnect size. For the typical rack size of 40 to 60 servers, both Scheme II and Scheme III can be used. Note that the curves of Scheme III show the case when the 2x2 coupler is used in the first stage. A bigger size of the coupler in the first stage increases the power budget and hence degrades the scalability.

If higher baud rate M-PAM is used in the aforementioned POI architectures, higher bandwidth receiver should be deployed to detect the signal. Consequently, the total noise which is proportional to the bandwidth of the photodiode would degrade the sensitivity of the receiver. Moreover, the unideal frequency response of the filter may have a negative impact on the BER performance. With the increase of the baud rate, the accumulated filtering effect of the link may distort the signal at the higher band part and thus degrade the sensitivity. As a result, the maximum size of the POI would be inferior to the numerical results shown in Figs. 9 and 10. On the other hand, the maximum data rate that can be supported at each size of the interconnect, would be higher.

Noted that the proposed methodology can also applied to quantify the scalability of POI where other modulation techniques, e.g., discrete multi-tone (DMT), are used. The upper bound of the scalability of such POI architectures can be obtained through the comparison between the link loss and system power budget which can be obtained by theoretical analysis of the transmission performance [36], [37], simulation [38], and experiment works [39].

IV. Conclusions

We propose a methodology for the scalability evaluation of optical interconnects in DCNs. To the best of our knowledge, the proposed methodology is the first taking into consideration different physical-layer features such as insertion loss, receiver noise, modulation format, symbol spacing, and decision thresholds. The methodology offers a comprehensive instrument for joint optimization of interconnect architecture and transmission parameters, while ensuring that BER requirements are fulfilled. By applying the proposed methodology for analyzing coupler-based POI architectures, we found that the quadratically spaced M-PAM is optimum when APD is employed, while receivers with PIN diodes achieve the best BER performance using equispaced modulation and decision thresholds. In a case study, two coupler-based POI architectures are identified that are able to offer the port count larger than 500.

Acknowledgments

This research was supported by a scholarship from the China Scholarship Council (CSC) under the Grant CSC No. 201406160028, the Swedish Research Council (VR), the Swedish Foundation for Strategic Research (SSF), the Göran Gustafsson Foundation and the National Natural Science Foundation of China (Grant No. 61550110240, 61671212).

References

[1] "Metro network traffic growth: an architecture impact study," Bell Labs, 2013 [Online]. Available: http://www.tmcnet.com/tmc/whitepapers/documents/whitepapers/2013/9378-bell-labs-metro-network-traffic-growth-an-architecture.pdf

[2] "The Zettabyte era-trends and analysis," Cisco white paper, 2014 [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/s-ervice-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html.

[3] C. Kachris and I. Tomkos, "A survey on optical interconnects for data centers," *IEEE Commun. Surv. Tut.*, vol. 14, no. 4, pp. 1021–1036, 2012.

[4] C. Kachris, K. Bergman, and I. Tomkos (Eds.), Optical Interconnects for Future Data Center Networks, Springer, 2013.

[5] L. A. Barroso, J. Clidaras and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines." *Synthesis lectures on computer architecture*, vol. 8, no. 3, pp. 1–154, 2013.

[6] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: a hybrid electrical/optical switch architecture for modular data centers," *ACM SIGCOMM Computer Communication Rev.*, vol. 41, no. 4, pp. 339–350, 2010.

[7] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. S. E. Ng, M. Kozuch, and M. Ryan, "c-Through: part-time optics in data centers," *ACM SIGCOMM Computer Commun. Rev.*, vol. 40, no. 4, pp. 327–338, 2010.

[8] Y. Yin, R. Proietti, X. Ye, C. J. Nitta, V. Akella, and S. J. B. Yoo, "Lions: An AWGR-based low-latency optical switch for high-performance computing and datacenters," *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 2, paper 3 600 409, 2013.

[9] S. Yan, E. Hugues-Salas, V. J. F. Rancaňo, Y. Shu, G. M. Saridis, B. R. Rofoee, Y. Yan, A. Peters, S. Jain, T. May-Smith, P. Petropoulos, D. J. Richardson, G. Zervas, and D. Simeonidou, "Archon: A function programmable optical interconnect architecture for transparent intra and inter data center SDM/TDM/WDM networking," *J. Lightw. Technol.*, vol. 33, no. 8, pp. 1586–1595, 2014.

[10] K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, and Y. Chen, "OSA: an optical switching architecture for data center network switch unprecedented flexibility," *IEEE/ACM Trans. Netw.*, vol. 22, no. 2, pp. 498–511, 2014.

[11] M. Fiorani, S. Aleksic, M. Casoni, L. Wosinska, and J. Chen, "Energy-efficient elastic optical interconnect architecture for data centers," *IEEE Commun. Lett.*, vol. 18, no. 9, pp. 1531–1534, 2014.

[12] Y. Gong, X. Hong, Y. Lu, S. He, and J. Chen, "POIs at top of the rack: offering high energy efficiency for datacenters," *Opt. Exp.*, vol. 23, no. 6, pp. 7957–7970, 2015.

[13] W. Ni, C. Huang, Y. L. Liu, W. Li, K.W. Leong, and J. Wu, "POXN: A new passive optical cross-connection network for low-cost power-efficient datacenters," *J. Lightw. Technol.*, vol. 32, no. 8, pp. 1482–1500, 2014.

[14] F. Effenberger *et al*. "An introduction to PON technologies." in *IEEE Commun. Mag.*, vol. 45, no. 3, pp. 17–25, 2007

[15] J. Chen, Y. Gong, M. Fiorani, and S. Aleksic, "Optical interconnects at top of the rack for energy-efficient datacenters," *IEEE Commun. Mag.*, vol. 53, no. 8, pp. 140–148, 2015.

[16] R. Pries, M. Jarschel, D. Schlosser, M. Klopf, and P. Tran-Gia, "Power consumption analysis of data center architectures," *in Proc. International Conference on Green Communications and Networking*, pp. 114–124, 2012.

[17] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement," *in Proc. INFOCOM*, 2010.

[18] A. S.-W. Tam, K. Xi, and H. J. Chao, "Use of devolved controllers in data center networks," in Proc. INFOCOM WKSHPS, pp. 596–601, 2011.

[19] K. Szczerba, P. Westbergh, J. Karout, J. S. Gustavsson, Å. Haglund, M. Karlsson, P. A. Andrekson, E. Agrell, and A. Larsson, "4-PAM for high-speed short-range optical communications," *IEEE J. Opt. Commun. Netw.,* vol. 4, no. 11, pp. 885–894, 2012.

[20] A. Bechtolsheim and M. Paniccia, "100G CLR4 Industry Alliance," 2014 [Online]. Available: http://www.clr4-alliance.org/media/doc/100G_CLR 4_Press_Deck.pdf

[21] Y. Cheng, M. Fiorani, L. Wosinska, and J Chen, "Reliable and cost efficient POIs for data centers", *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 1913–1916, 2015.

[22] C. K. Chan, K. L. Sherman, and M. Zirngibl, "A fast 100-channel wavelength-tunable transmitter for optical packet switching," *IEEE Photon. Technol. Lett.*, vol. 13, no. 7, pp. 729–731, 2001.

[23] J. E. Simsarian, P. Bernasconi, J. Gripp, M. C. Larson, D. T. Neilson, and J. Sinsky, "Fast tunable lasers for optical routers and networks," *in Proc. CLEO*, 2006.

[24] J. Gripp, J. E. Simsarian, J. D. Legrange, P. G. Bernasconi, and D. T. Neilson, “Architectures, components, and subsystems for future optical packet switches,” *IEEE J. on Sel. Topics in Quantum Electron.*, vol. 16, no. 5, pp. 1394–1404, 2010.

[25] A. Iocco, H. G. Limberger, R. P. Salathe, L. A. Everall, K. E. Chisholm, J. A. Williams, and I. Bennion, “Bragg grating fast tunable filter for wavelength division multiplexing”. *J. Lightw. Technol.* vol. 17, no.7, pp.1217–104, 1999.

[26] L. Kazovsky, S. Benedetto, and A. Willner, Optical fiber communication systems, Artech House, 1996.

[27] J. C. Palais, Fiber optic communications, Prentice Hall, 1988.

[28] S. Sumita and Y. Kokubun, "Low-loss wavelength selective switch using Mach-Zehnder coupled microring resonator," *in Proc. Optoelectronics and Commun. Conf. (OECC)*, 2010,

[29] “Fiber optic in-line isolators,” Newport datasheet, [Online]. Available: http://www.nxtbook.com/nxtbooks/newportcorp/resource20-11/#/326

[30] “1310/1550 nm high power optical circulator,” Agiltron datasheet [Online]. Available: http://www.agiltron.com/PDFs/1310-1550%20high%20power%20circulator.pdf

[31] R. S. Guzzon, E. J. Norberg, J. S. Parker, L. A. Johansson, and L. A. Coldren, “Integrated InP-InGaAsP tunable coupled ring optical bandpass filters with zero insertion loss,” *Opt. Exp.*, vol. 19, no. 8, pp. 7816–7826, 2011.

[32] G. P. Agrawal, Fiber-optic communication systems. John Wiley & Sons，2002.

[33] J. G. Proakis and M. Salehi, Digital Communications. 5th ed, McGraw-Hill Education, 2007.

[34] J. Rebola and A. Cartaxo, “On the quaternary level spacing signaling optimisation for increasing the transmission distance in optical communication systems”, *in Proc. Conf. on Telecommunications,* vol. 1, pp. 514–518, 2001.

[35] “10 Gbps tunable DWDM 80 km SFP+ optical transceiver,” Brocade datasheet, 2013 [Online]. Available: http://www.brocade.com/content/d-am/common/documents/content-types/datasheet/10gbe-tunable-dwdm-80km-sfp-ds.pdf

[36] P. K. Vitthaladevuni, M. S. Alouini, and J. C. Kieffer, “Exact BER computation for cross QAM constellations,” *IEEE Trans. Wirel. Commun.*, vol. 4, no. 6, pp. 3039–3050, 2005.

[37] M. H. and S. H. L. Peng, “On Bit-Loading for Discrete Multi-Tone Transmission Over Short Range POF Systems,” *J. Light. Technol.*, vol. 31, no. 24, pp. 4155–4165, 2013.

[38] R. Lin, K. Szczerba, E. Agrell, L. Wosinska, M. Tang, and J. Chen, “To overcome the scalability limitation of passive optical interconnects in datacentres,” *in Proc. Asia Communications and Photonics Conference*, ATh3E.3, 2016.

[39] R. Lin, X. Pang, O. Ozolins, Z. Feng, A. Djupsjöbacka, and U. Westergren, “Experimental Validation of Scalability Improvement for Passive Optical Interconnect by Implementing Digital Equalization,” *in Proc. European Conf. on Opt. Commun., ECOC*, 2016.